UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

RICHARD KADREY, et al.,

Plaintiffs,

v.

META PLATFORMS, INC.,

Defendant.

Case No. 23-cv-03417-VC (TSH)

DISCOVERY ORDER

Re: Dkt. No. 267

A. ECF No. 267

1. Issue #4

a. RFPs 64, 77, 45, 46, 53, 54, 59

Plaintiffs move to compel on seven RFPs.

RFP 64: "Documents and Communications sufficient to show each instance within the last three years where You have licensed copyrighted works for Meta's commercial use."

The Court agrees with Meta that this RFP is unreasonably overbroad because it seeks information concerning each instance in which Meta licensed a copyrighted work for Meta's commercial use, regardless of whether the commercial use had anything to do with AI or any issue that is relevant to this case. This would include, for example, licensing a song to use in an advertisement. Plaintiffs' motion is **DENIED** as to RFP 64.

RFP 77: "Communications Concerning any licensing copyrighted works that were used to train the Meta Language Models."

The Court reads this RFP as if there were an "of" between "licensing" and "copyrighted." Meta argues that with respect to copyrighted textual works, no such licenses exist and thus there is nothing to produce. However, this RFP is not limited to textual works, and although Plaintiffs'

copyrighted works are textual in nature, Meta does not explain why relevant evidence would be limited to textual works. The Court also does not think that "communications concerning" any licensing of copyrighted works are limited to communications that successfully resulted in a license. Communications would also be responsive if they resulted in no license being obtained. The Court does agree with Meta that the words "that were used to train the Meta Language Models" means the RFP is limited to communications concerning any licensing of copyrighted works that were, in fact, used to train the Meta Language Models. Accordingly, the Court **GRANTS** Plaintiffs' motion as to RFP 77 in part and **ORDERS** Meta to produce responsive documents regardless of whether the communications successfully resulted in a license, and not limited to textual works.

RFP 45: "All Documents and Communications Concerning any licensing, accreditation, or attribution mechanism, or similar tool for crediting, compensating, or seeking consent from owners of copyrighted works that were used to train the Meta Language Models."

Meta does not dispute the relevance of this information but argues it has no responsive documents. Accordingly, the Court **GRANTS** Plaintiffs' motion as to RFP 45. If there really is nothing responsive, then there is nothing for Meta to produce.

RFP 46: "All Documents and Communications sufficient to show Your actual or projected income from the sale or licensing of the Meta Language Models." RFP 53: "All Documents and Communications Concerning any income statement, balance sheet, or statement of cash flows, Concerning any of the Meta Language Models."

Meta does not dispute the relevance of this information and states it is not withholding documents responsive to these RFPs. Accordingly, the Court **GRANTS** Plaintiffs' motion as to RFPs 46 and 53. If Meta has already produced the responsive documents, then there is nothing more for Meta to produce.

RFP 54: "All Documents and Communications Concerning any decision by You to not develop an interface for end users to interact with any of the Meta Language Models."

Other than asserting that the documents sought by RFP 54 are "clearly relevant," Plaintiffs have not actually explained how responsive documents are relevant to any issue in this case.

Plaintiffs' motion is **DENIED** as to RFP 54.

RFP 59: "Documents and Communications Concerning the ability of any Meta Language Model to output fictional works."

Plaintiffs do not explain why documents responsive to this RFP are relevant to this case. Their motion is **DENIED** as to RFP 59.

b. Time Frame for Document Productions

The parties agree that Meta has limited the relevant time period for document production to January 1, 2022 to the present. Plaintiffs argue that this time frame is too narrow. They say that the proposed class period begins on July 7, 2020 (three years before the Complaint was filed). They also argue that copying or discussions "may have" occurred before 2022. They say "[t]he Court should require Meta to run searches (and produce documents from) as far back as necessary to capture all instances in which Meta copied—or discussed copying—copyrighted data that was used to train Llama, or, at a minimum, as far back as the beginning of the class period."

The request for Meta to produce documents from "as far back as necessary" to capture relevant conduct is indeterminate. The Court has to specify a time frame. The Court therefore considers the alternative request to expand the time frame to "the beginning of the class period." That is a request to expand document production by a year and a half. This is the sort of request the Court does not expect to be made on the last day to move to compel concerning existing written discovery, which is when Plaintiffs filed this request. The date range for document production is something that should be resolved much sooner than that. Plaintiffs have filed a motion to compel 35 days before the close of fact discovery seeking an additional year and a half of document production. That likely cannot be done by the close of fact discovery, and ordering Meta to try threatens to turn the close of fact discovery into a train wreck. The Court continues to be concerned by Plaintiffs' repeated attempts to seek major expansions of the scope of discovery right near the end of fact discovery. A fire drill in the last month of fact discovery concerning a foundational issue that could have been raised much sooner is not proportional to the needs of the case. Plaintiffs' motion to expand the time frame for document production is **DENIED**.

2. Issue #5

a. Llama Source Code

Plaintiffs move to compel additional Llama source code. Meta argues that Plaintiffs have no existing written discovery requests that seek source code, and that such requests were first served on October 9, 2024, and Meta's responses were due on November 8, 2024, which is the date the joint discovery letter brief was filed.

Plaintiffs' section of the letter brief does not identify any discovery request to which source code is responsive. A party moving to compel should demonstrate that it asked for the materials in discovery. Plaintiffs have not shown that. Accordingly, their motion to compel is **DENIED** as to the source code.

b. Llama Training Data

Plaintiffs move to compel on RFPs 1-3 and 7 and rog 1. RFPs 1-3 seek "[t]he Training Data" for Llama 1, 2 and 3. RFP 7 seeks "[d]ocuments and Communications to, from, or with Library Genesis (aka LibGen) Concerning Training Data." Rog 1 asks:

> Describe in detail the data You have used to train or otherwise develop the Meta Language Models, Including, for each:
>
> a. How You obtained the data, e.g., by scraping the data, purchasing it from third parties, or by other means;
>
> b. All sources of Data, including any third parties that provided data sets;
>
> c. To the extent the data was derived from publicly available websites, a list of all such websites and, for each, the percentage of the data corpus that is derived from that website;
>
> d. The categories of content included in the data and the extent to which each category is represented in the data corpus (i.e., as a percentage of data used to train the model);
>
> e. All policies and procedures Related to identifying, assessing, vetting and selecting sources of data for the model.

Plaintiffs do not present any argument with respect to RFP 7. Meta states that it is not aware of any responsive documents. The Court **DENIES** the motion to compel as to RFP 7 because the motion is unexplained.

With respect to the training data, Plaintiffs say they need more information about how

Meta obtained and used it. Plaintiffs say that if the only issue in the case were the important binary question of whether Plaintiffs' copyrighted materials were in Meta's possession in some fashion, the data already produced would answer that question. But Plaintiffs say it is also important that they be permitted discovery that goes to the importance of and breadth of use of the copyrighted protected materials at issue. For these reasons, Plaintiffs do not believe it is enough for them to have access to the set of training data for Llamas 1-3, and submit they are entitled to information from Meta that identifies the iterations of copies of training data with copyrighted material or books within their possession, custody or control. Plaintiffs seek information on how many times Meta downloaded each copyrighted work, from where it downloaded each, when it downloaded each, and how it is using each copy. To be clear, in light of Meta's stated burden objection, Plaintiffs are not demanding that Meta produce each iteration of the copies. They would settle for a declaration or an amended answer to rog 1 that provides this information.

Meta has several responses. The major one is that Meta has identified and produced copies of the actual book-related training datasets that allegedly include copyrighted works that were actually used to train the Llama models. Meta argues that Plaintiffs' request that Meta scour the entire company to determine if there are other stored and duplicative copies of those datasets – copies which would not have been the ones used to train Llama – is overly burdensome and not proportional to the needs of the case.

The Court **DENIES** Plaintiffs' motion to compel because RFPs 1-3 and rog 1 (and RFP 7) did not ask for this information. RFPs 1-3 asked for the training data for Llama 1, 2 and 3. If there are other copies of the same copyrighted works in Meta's possession that were not used to train Llama, the RFPs didn't ask for those. Similarly, rog 1 asked Meta to "[d]escribe in detail the data You have used to train or otherwise develop the Meta Language Models . . ." This rog didn't ask about the full scope of what Meta has or does with the copyrighted works. It just asked about the data used to train or otherwise develop the language models. Plaintiffs do not make any argument that Meta's response to rog 1 is incomplete or otherwise defective. Plaintiffs argue that they seek "information about how books were used in LLM training and operationalization, and how in turn the LLMs or book corpuses are used by Meta." Rog 1 did not ask for that. It asked

Meta to describe the data it used to train or develop the models, including (a) how Meta got it, (b) all the sources of data, (c) the public websites it got data from, (d) the categories of content included in the data, and (e) the policies and procedures for selecting sources of data. The rog asked for what data Meta used and where and how and why it got it. It did not ask anything about how the data was used in training or operationalization, or how the LLMs or book corpuses are used by Meta. Accordingly, Plaintiffs' motion is **DENIED** as to these discovery requests.

IT IS SO ORDERED.

Dated: November 25, 2024

THOMAS S. HIXSON
United States Magistrate Judge